



Estadística

La Estadística tiene sus antecedentes históricos en los famosos censos, que consistían en observaciones sistemáticas y periódicas sobre datos de la población para fines de guerra y finanzas realizados desde antes de Cristo.

Si bien no hay una definición de *estadística* exacta, se puede decir que la “estadística es la ciencia que trata de la recolección, presentación, análisis y uso de datos para la toma de decisiones”.

La Estadística nos ofrece dos tipos de investigación:

- **Estadística descriptiva:** puede definirse como aquellos métodos que incluyen la recolección, presentación y caracterización de un conjunto de datos con el fin de describir apropiadamente las diversas características de ese conjunto de datos.
- **Estadística inferencial:** puede definirse como aquellos métodos que hacen posible la estimación de una característica de una población o la toma de una decisión referente a una población, basándose sólo en los resultados de una muestra.

Sus aplicaciones se dan en todos los campos de la investigación, siendo utilizada como medio auxiliar entre otros por economistas, médicos, físicos y técnicos. Además, ya se hace necesario tener ciertos conocimientos elementales de Estadística para la lectura de un diario o recoger información de un noticiero.

El método Estadístico

Primera etapa: recuento o recopilación de datos

Para hacer el estudio de los caracteres estadísticos que se desea conocer se hacen observaciones. Si las observaciones se realizan sobre el total de un grupo, a ese grupo se lo denomina **población** o **universo** y cada observación se llama **individuo**.

Si se efectúa un censo ganadero:

Población —————→ el conjunto de animales

Individuo —————→ cada animal

Si se efectúa un censo sobre viviendas:

Población —————→ el total de viviendas

Individuo —————→ cada vivienda

Muchas veces es muy difícil estudiar el número total (N) de los individuos de una población debido al costo económico, el tiempo que ocasiona o por ser muy difícil de delimitar ese número total.



Por ejemplo: si se desea saber la diferencia de presión sanguínea entre hombres negros y blancos, sería imposible determinar la presión de todos los hombres negros y de todos los hombres blancos. El problema se resuelve recurriendo a las **muestras**.

Se llama **muestra** al conjunto de n individuos ($n < N$) elegidos al azar entre los N de una población dada.

Luego si en el ejemplo anterior se toma la presión a 200 hombres elegidos al azar:

Población —————→ total de hombres existentes

Muestra —————→ los 200 hombres elegidos

Individuo —————→ cada hombre

Para que los estudios realizados sobre la muestra sean válidos, la muestra debe ser representativa de la población.

Si una medida de resumen se calcula para describir una característica de toda una población, ésta es denominada **parámetro**. En cambio, si se calcula para describir una característica de una sola muestra de la población, la misma recibe el nombre de **estadístico**.

Ejemplo:

A una consultora le encargan hacer un estudio acerca del cuál es la intención del voto de los ciudadanos de una ciudad en las próximas elecciones. Como no es posible encuestar a todos los ciudadanos, la consultora toma un grupo de 500 y sobre él analiza la elección de cada uno. Con los datos recopilados sobre esta muestra se puede hacer una proyección de los votos que obtendrá cada candidato.

El tema que es objeto de estudio en una población determinada es la **variable**. En el ejemplo anterior, la variable analizada es la elección que efectúa cada ciudadano.

Cuando las variables se expresan mediante una cantidad, como el peso, la altura, etc, son **cuantitativas**. Este tipo de variable puede ser de dos tipos:

- **Variable cuantitativa discreta:** es aquella que asume un conjunto de valores que están en correspondencia biunívoca con los números naturales o un subconjunto de estos.
- **Variable cuantitativa continua:** es aquella que asume un conjunto de valores que están en correspondencia con los números reales, siendo una cantidad no enumerable o infinita de valores dentro de un intervalo.

En cambio si indican una cualidad o característica de la población, por ejemplo el estado civil, el sexo, etc, son **cualitativas**. A las variables cualitativas también se las llama **atributos**.

Segunda etapa: tabulación y graficación.

Cuando se realiza el censo o relevamiento, los datos de cada individuo se anotan, generalmente, en una ficha o una planilla según la cantidad de datos requeridos.



Una vez recogidos los datos se pueden escribir en una tabla en forma ordenada. Esto constituye una *serie simple*.

Se considera, por ejemplo, las tallas de 40 alumnos de 3° año de Polimodal, ordenadas de menor a mayor y expresadas en centímetros.

N°	Talla	N°	Talla
1	150	21	163
2	150	22	164
3	154	23	164
4	155	24	164
5	156	25	164
6	156	26	164
7	157	27	165
8	157	28	165
9	158	29	165
10	158	30	166
11	158	31	166
12	159	32	167
13	159	33	167
14	160	34	167
15	160	35	168
16	160	36	169
17	160	37	169
18	160	38	170
19	161	39	170
20	163	40	171

Cuando se recopilan muchos datos, puede ser que algunos se repitan. Se llama *frecuencia* a la cantidad de veces que se repite un determinado valor de la variable.

La *frecuencia relativa* es que parte del total representa cada valor de la variable. Si se multiplica la frecuencia relativa por 100, se obtiene la *frecuencia relativa porcentual*.

Muchas veces interesa conocer cuántos datos se acumulan hasta un cierto valor, para lo cual habrá que sumar a la frecuencia de ese valor, la frecuencia de los valores anteriores. A esta suma parcial se la llama *frecuencia acumulada*.

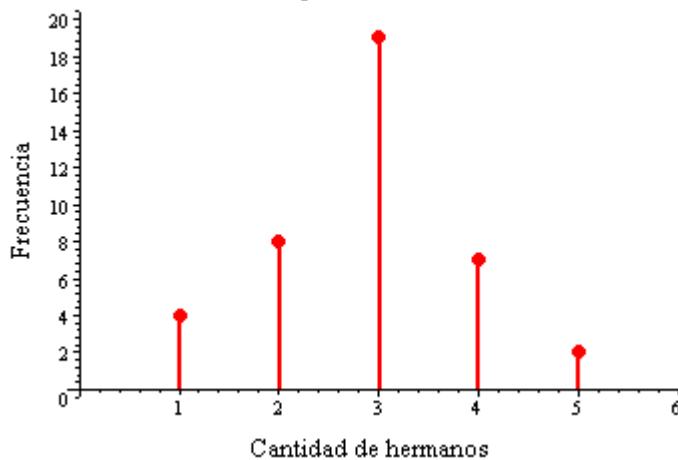
Ejemplo:

Se realizó una encuesta a 40 chicos acerca de la cantidad de hermanos que tiene cada uno. A continuación se mostrará la tabla que recibe el nombre de *distribución de frecuencias*, en la que figuran los datos recopilados:



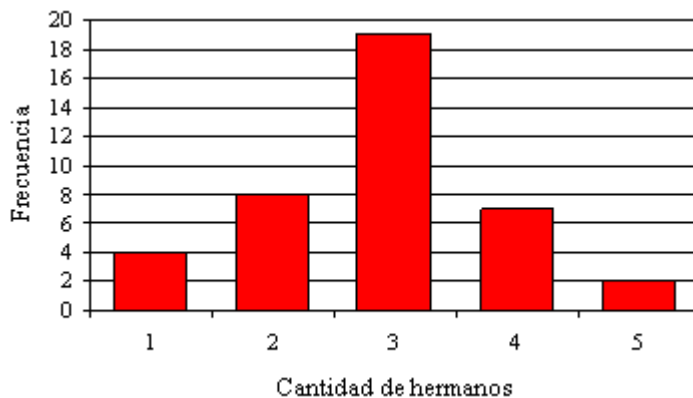
<i>Cantidad de hermanos</i>	<i>Frecuencia</i>	<i>Frecuencia relativa</i>	<i>Frecuencia acumulada</i>
1	4	0,1	4
2	8	0,2	12
3	19	0,475	31
4	7	0,175	38
5	2	0,05	40
Total	40	1	

Gráfico de bastones

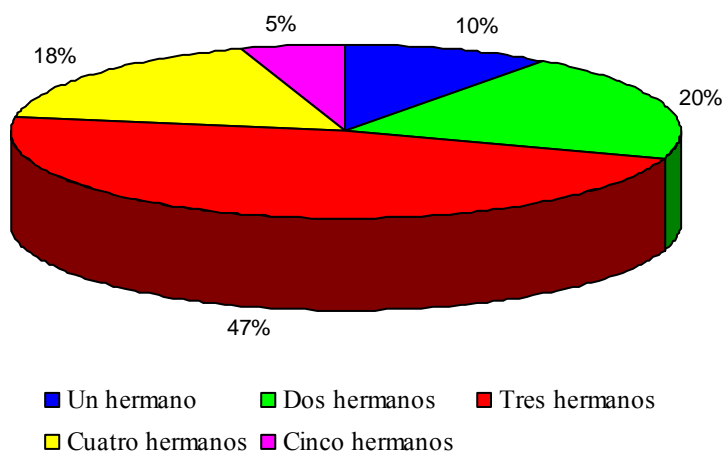


Aunque las tablas bien construidas presentan claros los resultados y ofrecen un buen medio para poder obtener importantes conclusiones, no cabe duda de que si se las expresa mediante gráficos, éstos constituyen por sí mismos una poderosa herramienta para el análisis de los datos, siendo en ocasiones el medio más efectivo no sólo para describir y resumir la información, sino también para analizarla. Entre los gráficos más utilizados se pueden destacar:

- **Gráfico de barras:** es una de las formas más simples de representación. La longitud de cada barra es igual a la frecuencia de cada observación.



- **Gráfico circular:** el área de cada sector, representa el porcentaje correspondiente a la frecuencia de un valor de la variable. Es conveniente su utilización cuando el número de sectores es pequeño y sus áreas están bien diferenciadas.



- **Histograma:** se utiliza para representar una tabla de frecuencias de intervalos de clase. Para construir un gráfico de este tipo, se divide el rango de valores de la variable en intervalos de igual amplitud, representando sobre cada intervalo un rectángulo que tiene a este segmento como base. El criterio para calcular la altura de cada rectángulo es el de mantener la proporcionalidad entre las frecuencias absolutas de los datos en cada intervalo y el área de los rectángulos. Uniendo los puntos medios del extremo superior de las barras del histograma, se obtiene una imagen que se llama **polígono de frecuencias**. Dicha figura pretende mostrar, de la forma más simple, en qué rango se encuentra la mayor parte de los datos.

Si se considera nuevamente las tallas de los alumnos de 3° año de Polimodal, es posible determinar la cantidad de intervalos de clase utilizando la expresión:

$$\text{Número de intervalos de clase} \approx 1 + 3,3 \cdot \log(N)$$

donde N es la cantidad total de individuos que constituyen a la población. De esta manera:

$$\text{Número de intervalos de clase} \approx 1 + 3,3 \cdot \log(40) \approx 7$$

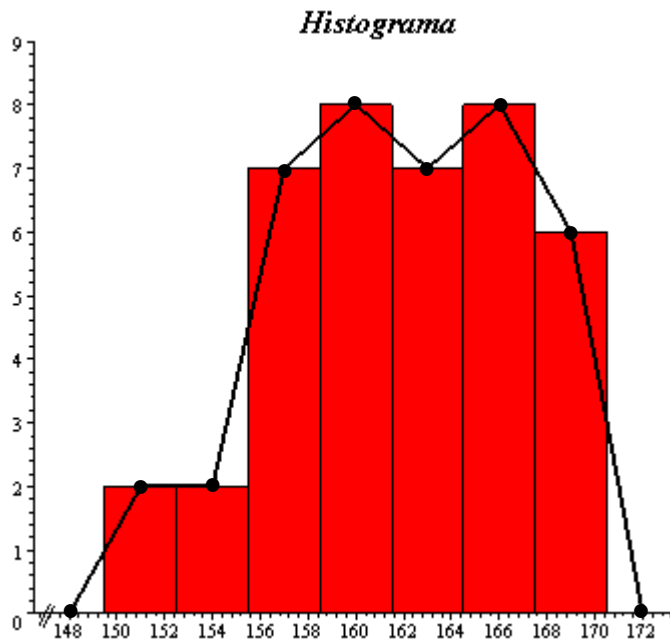
Para poder calcular la amplitud de cada uno de estos intervalos, basta con efectuar la división $[(b - a) + 1] \div \text{número de intervalos}$, donde a es el mínimo valor que toma la variable que se está analizando y b , el mayor.

$$\text{Amplitud} \approx [(171 - 150) + 1] \div 7 \approx 3$$

Para evitar que los extremos de los intervalos coincidan con algún valor de la variable, se los define restando media unidad al último decimal.



Talla	Frecuencia	Frecuencia relativa	Frecuencia acumulada
149,5 – 152,5	2	0,05	5
152,5 – 155,5	2	0,05	5
155,5 – 158,5	7	0,175	17,5
158,5 – 161,5	8	0,2	20
161,5 – 164,5	7	0,175	17,5
164,5 – 167,5	8	0,2	20
167,5 – 170,5	6	0,15	15
Total	40	1	



Tercera etapa: análisis y medición de los datos.

Una característica constante a lo largo de la Estadística es el manejo de una gran cantidad de datos. Uno de los fines importantes de la Estadística descriptiva es el de resumir o sintetizar esas grandes cantidades de datos en unos pocos números que nos reflejen de la forma más aproximada posible, el comportamiento de todos los elementos de una población con respecto al carácter que se desea estudiar. Estos números, como ya se mencionó anteriormente, se conocen con el nombre de **parámetros** si los mismos se calculan teniendo en cuenta la totalidad de la población.

Parámetros centrales

Los expertos en baloncesto quieren hacer estudios comparativos sobre las estaturas de los jugadores que componen los equipos de Primera División.



Las estaturas de los jugadores de los equipos A y B se resumen en las siguientes tablas:

<i>Equipo A</i>	
<i>Nº</i>	<i>Estatura</i>
1	183
2	184
3	187
4	188
5	189
6	189
7	193
8	196
9	199
10	202
11	203
12	204
13	206
14	207

<i>Equipo B</i>	
<i>Nº</i>	<i>Estatura</i>
1	181
2	186
3	187
4	191
5	191
6	192
7	192
8	192
9	196
10	197
11	197
12	201
13	206
14	207

Comparar las alturas de los dos equipos observando las tablas sería muy difícil. Por ello, para efectuar esta comparación, de manera sencilla, se calcula para cada equipo, un valor numérico que represente a las estaturas de sus jugadores, llamado **parámetro central**.

Los **parámetros centrales** son unos números que tienen como objetivo agrupar o centralizar los datos correspondientes a toda la población en un solo valor numérico, representante del conjunto total. Puede parecer demasiado simplista querer representar todos los datos de una población mediante un único número. Sin embargo, si bien no son suficientes, estos parámetros centrales o promedios son de gran utilidad para el manejo de datos estadísticos.

Existen varios promedios pero el uso ha impuesto unos pocos entre los que se destaca la **media aritmética**, la **mediana** y el **modo**.

Media aritmética

Es el parámetro central que se utiliza con mayor frecuencia. Se llama **media aritmética** al cociente que se obtiene de dividir la suma de todos los valores por el número de éstos.

Se calculará la altura media del equipo A, recordando que se debe sumar todas las alturas y dividir por el número de jugadores.

$$\bar{x} = \frac{183 + 184 + 187 + 188 + 189 + 189 + 193 + 196 + 199 + 202 + 203 + 204 + 206 + 207}{14}$$

$$\bar{x} = \frac{2730}{14} = 195 \text{ cm}$$



Del mismo modo, se calcula la altura promedio del equipo B.

$$\bar{x} = \frac{181 + 186 + 187 + 191 + 191 + 192 + 192 + 192 + 196 + 197 + 197 + 201 + 206 + 207}{14}$$

$$\bar{x} = \frac{2716}{14} = 194 \text{ cm}$$

En general, si se llama x_1, x_2, \dots, x_n a los n valores que toma la variable, se designa por \bar{x} a la media aritmética y se calcula así:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Cálculo de la media en distribuciones con datos agrupados

Las notas obtenidas por los alumnos de un curso se expresan mediante la siguiente tabla y su correspondiente gráfica:

Notas	N° de alumnos
1	1
2	2
3	3
4	6
5	9
6	4
7	2
8	2
9	2
10	1

La media aritmética se calcularía así:

$$\bar{x} = \frac{1 + (2 + 2) + (3 + 3 + 3) + (4 + 4 + 4 + 4 + 4 + 4) + \dots + (8 + 8) + (9 + 9) + 10}{1 + 2 + 3 + 6 + \dots + 2 + 2 + 1}$$

El resultado de cada paréntesis del numerador es igual al producto de cada nota por su correspondiente frecuencia. El denominador es la suma de las frecuencias.

Por lo tanto:

$$\bar{x} = \frac{1 + 2 \times 2 + 3 \times 3 + 4 \times 6 + 5 \times 9 + 6 \times 4 + 7 \times 2 + 8 \times 2 + 9 \times 2 + 10}{32} = \frac{165}{32} \approx 5,17$$



Notas (x_i)	Frecuencia (f_i)	$x_i \cdot f_i$
1	1	1
2	2	4
3	3	9
4	6	24
5	9	45
6	4	24
7	2	14
8	2	16
9	2	18
10	1	10
Total	32	165

En general, para una distribución que se da con los datos agrupados, la media aritmética es:

$$\bar{x} = \frac{\sum_{i=1}^n x_i \cdot f_i}{\sum_{i=1}^n f_i}$$

Cálculo de la media aritmética en distribuciones con datos agrupados en intervalos

¿Cómo se calcula la media de las estaturas del equipo A si sólo se dispone de la siguiente tabla?

Equipo A	
Estaturas	Nº de jugadores
180,5 – 185,5	2
185,5 – 190,5	4
190,5 – 195,5	1
195,5 – 200,5	2
200,5 – 205,5	3
205,5 – 210,5	2

En estos casos lo más razonable es asignarles a todos el valor central del intervalo en el que están. A los que se encuentran entre 180,5 y 185,5 se les asigna el valor 183; a los que están entre 185,5 y 190,5 se les asigna 188 y así sucesivamente.

Para calcular la media aritmética, se amplía la tabla con los valores centrales y la columna $P_{mi} \cdot f_i$.



Estaturas (x_i)	N° de jugadores (f_i)	P_{mi}	$P_{mi} \cdot f_i$
180,5 – 185,5	2	183	366
185,5 – 190,5	4	188	752
190,5 – 195,5	1	193	193
195,5 – 200,5	2	198	396
200,5 – 205,5	3	203	609
205,5 – 210,5	2	208	416
Total	14	–	2732

La media aritmética se calcula de la siguiente forma:

$$\bar{x} = \frac{\sum_{i=1}^n P_{mi} \cdot f_i}{\sum_{i=1}^n f_i} = \frac{2732}{14} \approx 195,14$$

Se observa que el valor obtenido 195,14 cm no coincide con el que se obtuvo manejando las estaturas exactas de cada jugador, 195 cm, pero se aproxima mucho.

Otros valores centrales interesantes, por su significado, son la **mediana** y la **moda**. No son propiamente promedios porque sus valores no dependen de los valores particulares sino de la densificación u ordenación de los mismos. No obstante, resultan de gran interés porque tienen un significado claro y expresivo, que los hace útiles al estudiar la estructura de la serie.

Mediana

La **mediana**, es un valor que separa la serie en dos partes de igual número de términos, de tal manera que en uno de los grupos queden términos inferiores a la mediana y en el otro, superiores. En el caso de datos no agrupados en intervalos, efectuando un ordenamiento de valores de la serie de menor a mayor, la mediana será el valor central de la ordenación, si el número de valores es impar y el promedio de los dos valores centrales si es par.

Ejemplo: Las estaturas de cinco alumnos son las siguientes:

1,47 m; 1,68 m; 1,53 m; 1,48 m; 1,70 m

Efectuando un ordenamiento de menor a mayor:

1,47 m; 1,48 m; 1,53 m; 1,68 m; 1,70 m

En este caso la mediana es 1,53 m.



Si en cambio, los datos están agrupados de la manera que se presentan a continuación, el cálculo de la mediana se realiza de la siguiente forma:

Al contar el número de letras que tienen 116 palabras de un artículo sobre Estadística, los resultados fueron:

Número de letras	Número de palabras
1	4
2	36
3	14
4	9
5	15
6	7
7	6
8	9
9	8
10	8

Para calcular la mediana, se ampliará la tabla con una columna en la que se escribirá la **frecuencia acumulada** de cada valor de la variable.

Número de letras (x_i)	Número de palabras (f_i)	Frecuencia acumulada (F)
1	4	4
2	36	40
3	14	54
4	9	63
5	15	78
6	7	85
7	6	91
8	9	100
9	8	108
10	8	116

La mediana, efectuando un ordenamiento de menor a mayor, es aquél valor de la variable que divide a los términos de la serie en dos partes. En una serie de frecuencias se consideran las frecuencias acumuladas. Cada frecuencia se obtiene sumando a la correspondiente las frecuencias anteriores. La mediana corresponde a la observación cuya frecuencia acumulada contiene a:

$$\frac{N}{2}$$

donde N es la cantidad total de individuos que constituyen a la población. En el ejemplo considerado, la mediana es igual a 4.

En cambio, si se dispone de la siguiente tabla, ¿cómo se calcula la mediana en esta situación?

Número de letras	Número de palabras
0,5 – 2,5	40
2,5 – 4,5	23
4,5 – 6,5	22
6,5 – 8,5	15
8,5 – 10,5	16

Para determinar la media en esta situación, se amplía la tabla con el cálculo de la frecuencia acumulada.

Número de letras (x_i)	Número de palabras (f_i)	Frecuencia acumulada (F)
0,5 – 2,5	40	40
2,5 – 4,5	23	63
4,5 – 6,5	22	85
6,5 – 8,5	15	100
8,5 – 10,5	16	116

La mediana se calcula de la siguiente forma:

$$M_{na} = L_i + \frac{\frac{N}{2} - F_{i-1}}{f_i} \cdot a$$

donde L_i es el límite inferior del intervalo que contiene la observación $N/2$ y f_i , su frecuencia, F_{i-1} es la frecuencia acumulada del intervalo anterior que contiene a la mediana y a es la amplitud del intervalo.

Para este caso particular:

$$M_{na} = 2,5 + \frac{58 - 40}{23} \cdot 2 \approx 4,06$$

Modo o moda

Se llama **modo** o **moda** al valor de la variable correspondiente a la máxima frecuencia. El modo se encuentra inmediatamente, al observar en la tabla el valor de la variable que se repite una mayor cantidad de veces.

El modo es un valor central expresivo que debe utilizarse con cierta precaución cuando no se destaca claramente del resto de la distribución.

Para explicar cómo se calcula el modo, se considerará las notas de un examen de inglés de una clase. La siguiente tabla muestra los resultados obtenidos.



Notas	Número de alumnos
1	4
2	3
3	2
4	1
5	2
6	3
7	8
8	3
9	2
10	1

En este caso, el modo de la distribución es el valor 7.

¿Cómo se calculará el modo si se tiene la siguiente tabla como información?

Notas	Número de alumnos
0,5 – 2,5	7
2,5 – 4,5	3
4,5 – 6,5	5
6,5 – 8,5	11
8,5 – 10,5	3

El modo se calcula de la siguiente forma:

$$M_0 = L_i + \frac{d_1}{d_1 + d_2} \cdot a$$

donde L_i es el límite inferior del intervalo que tiene mayor frecuencia y f_i , la frecuencia de ese intervalo, $d_1 = f_i - f_{i-1}$ (f_{i-1} es la frecuencia de intervalo anterior que contiene al modo) y $d_2 = f_i - f_{i+1}$ (f_{i+1} es la frecuencia de intervalo siguiente que contiene al modo).

Para este caso particular:

$$M_0 = 6,5 + \frac{11}{6 + 8} \cdot 2 \approx 8,07$$

Medidas de posición

La mediana resuelve el problema de dividir los términos de la serie, ordenados según su valor, en mitades. Se plantea ahora, la cuestión de dividir en cuatro grupos a la población. Los puntos de esta división se llaman **cuartiles**.



El primer cuartil, que se indica con Q_1 , es un valor tal que supera la cuarta parte de los valores de la serie, y es superado por las tres cuartas partes restantes. El segundo cuartil, Q_2 , coincide con la mediana. El tercer cuartil, Q_3 , supera en valor a las tres cuartas partes de los términos y es superado por la cuarta parte restante.

Los **deciles** y los **centiles** se definen de la misma forma que los cuartiles; los primeros dividen la población ordenada en diez partes y los segundos en cien.

Estas medidas tienen variadas aplicaciones. Por ejemplo, pueden utilizarse como medidas de dispersión, es decir, como medidas que indican la concentración de la población.

Gráficamente, los cuartiles, deciles y centiles pueden ser determinados mediante el polígono de frecuencias acumuladas, de manera análoga al caso de la mediana. En forma analítica, se utilizan fórmulas aproximadas. Así, si se quiere calcular el tercer cuartil en el ejemplo del número de letras que tienen 116 palabras de un artículo sobre Estadística, se procede de la siguiente manera:

Número de letras (x_i)	Número de palabras (f_i)	Frecuencia acumulada (F)
1	4	4
2	36	40
3	14	54
4	9	63
5	15	78
6	7	85
7	6	91
8	9	100
9	8	108
10	8	116

El tercer cuartil corresponde a la observación cuya frecuencia acumulada contiene a:

$$\frac{3}{4}N$$

donde N es la cantidad total de individuos que constituyen a la población. En el ejemplo considerado, el tercer cuartil es igual a 7.

En cambio, si se los datos están agrupados en intervalos, el tercer cuartil se calcula así:

Número de letras (x_i)	Número de palabras (f_i)	Frecuencia acumulada (F)
0,5 – 2,5	40	40
2,5 – 4,5	23	63
4,5 – 6,5	22	85
6,5 – 8,5	15	100
8,5 – 10,5	16	116

$$Q_3 = L_i + \frac{\frac{3}{4}N - F_{i-1}}{f_i} \cdot a = 6,5 + \frac{87 - 85}{100} \cdot 2 = 6,54$$

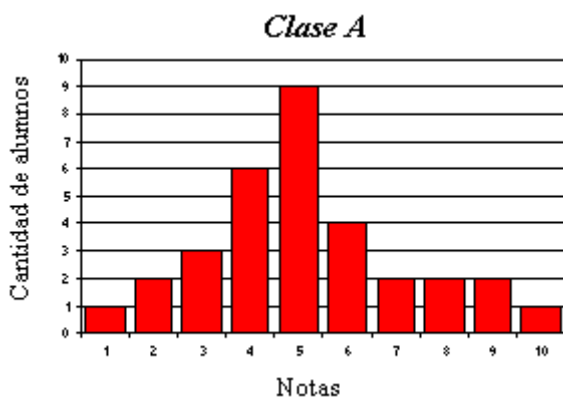
Medidas de dispersión

Un promedio resume todos los valores observados en uno solo que lo representa. La utilidad de un promedio depende, por lo tanto, de su poder representativo del conjunto de observaciones. Si los valores observados están muy concentrados alrededor del promedio, éste es muy representativo, pero si aquellos valores están muy dispersos con relación al promedio, éste es poco representativo. En consecuencia, el significado de un promedio gana mucho si viene acompañado de una medida de la concentración o dispersión de las observaciones en torno a él.

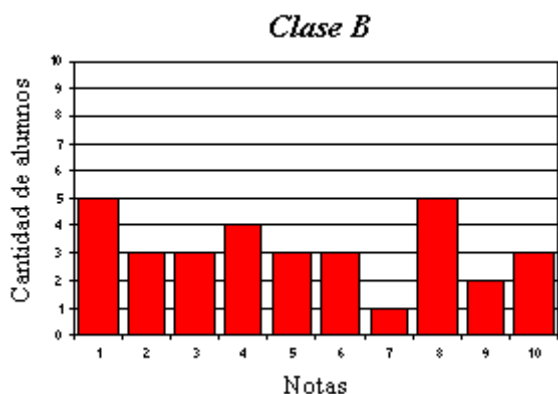
Entonces cuando se quiere conocer la **dispersión de una variable**, lo que se intenta es obtener un número, que indique el mayor o menor grado de variación o que nos proporcione una medida de la proximidad o lejanía de los datos respecto de su valor central.

Ejemplo:

Las notas obtenidas por los alumnos de dos clases se expresan mediante estas tablas y sus correspondientes gráficas:



Notas	Número de alumnos
1	1
2	2
3	3
4	6
5	9
6	4
7	2
8	2
9	2
10	1



Notas	Número de alumnos
1	5
2	3
3	3
4	4
5	3
6	3
7	1
8	5
9	2
10	3



¿Cuál es la nota media de cada clase?

$$\bar{x}_A = \frac{\sum_{i=1}^n f_i \cdot x_i}{\sum_{i=1}^n f_i} = \frac{165}{32} \approx 5,156 \qquad \bar{x}_B = \frac{\sum_{i=1}^n f_i \cdot x_i}{\sum_{i=1}^n f_i} = \frac{164}{32} \approx 5,125$$

Al calcular la nota media de cada clase, se obtiene unos valores muy parecidos, próximos a 5. Sin embargo, las distribuciones de las notas son muy distintas.

¿Qué significa esto? Que conociendo sólo la nota media de una clase, no se puede hacer una idea de cómo se distribuyen las calificaciones, es decir, no se sabe lo dispersas que están las notas con relación a la media.

Desviación

Se llama **desviación** del valor x_i respecto de su media a la diferencia entre dicho valor y su media, es decir: $d_i = x_i - \bar{x}$.

Esta diferencia da una medida de la proximidad de cada valor de la variable con respecto a la media aritmética. De acuerdo con la definición, estas desviaciones pueden ser positivas, negativas o nulas. La propiedad de la media aritmética que afirma que la suma de todas las desviaciones es igual a cero, hace que no se pueda utilizar esta suma para medir la dispersión. Para evitar esto se recurre a un procedimiento que da a lugar a un parámetro de dispersión llamado **variancia** y que consiste en trabajar con los cuadrados de dichas desviaciones. La variancia es una medida en la que no aparecen unidades originales porque las mismas figuran elevadas al cuadrado. Para solucionar este inconveniente se define la **desviación típica o standar**.

En el ejemplo considerado, el cálculo de la variancia y la desviación típica para la **clase A** se efectúa de la siguiente manera:

x_i	f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i \cdot (x_i - \bar{x})^2$
1	1	-4,156	17,2723	17,2723
2	2	-3,156	9,9603	19,9206
3	3	-2,156	4,6483	13,9449
4	6	-1,156	1,3363	8,0178
5	9	-0,156	0,0243	0,2187
6	4	0,844	0,7123	2,8492
7	2	1,844	3,4003	6,8006
8	2	2,844	8,0883	16,1766
9	2	3,844	14,7763	29,5526
10	1	4,844	23,4643	23,4643
Total	32	–	–	138,2176



Se suman todos los cuadrados por sus respectivas frecuencias y se divide entre el número de calificaciones:

$$\sigma^2 = \frac{\sum_{i=1}^n f_i \cdot (x_i - \bar{x})^2}{\sum_{i=1}^n f_i}$$

$$\sigma^2 = \frac{138,2176}{32} \approx 4,3193 \qquad \sigma = \sqrt{4,3193} \approx 2,078$$

De manera similar, se trabaja con la **clase B**.

x_i	f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i \cdot (x_i - \bar{x})^2$
1	5	-4,125	17,0156	85,078
2	3	-3,125	9,7656	29,2968
3	3	-2,125	4,5156	13,5468
4	4	-1,125	1,2656	5,0624
5	3	-0,125	0,0156	0,0468
6	3	0,875	0,7656	2,2968
7	1	1,875	3,5156	3,5156
8	5	2,875	8,2656	41,328
9	2	3,875	15,0156	30,0312
10	3	4,875	23,7656	71,2968
Total	32	-	-	281,4992

$$\sigma^2 = \frac{281,4992}{32} \approx 8,7969 \qquad \sigma = \sqrt{8,7969} \approx 2,9659$$

La desviación típica en la clase B es mayor que en la clase A, eso significa que las notas de los alumnos de la clase B están más separadas de la media que las de la clase A. Es decir, cuanto mayor es la desviación típica, más dispersos están los datos respecto de la media.

Cálculo de la desviación típica en distribuciones con datos agrupados en intervalos

¿Cómo calculamos la desviación típica o estándar de la **clase A** si los datos ahora están agrupados en intervalos?

Notas	Número de alumnos
0,5 – 2,5	3
2,5 – 4,5	9
4,5 – 6,5	13
6,5 – 8,5	4
8,5 – 10,5	3

En primer lugar, se debe determinar la media aritmética de la población.

x_i	f_i	P_{mi}	$P_{mi} \cdot f_i$
0,5 – 2,5	3	1,5	4,5
2,5 – 4,5	9	3,5	31,5
4,5 – 6,5	13	5,5	71,5
6,5 – 8,5	4	7,5	30
8,5 – 10,5	3	9,5	28,5
Total	32	–	166

En este caso, la media aritmética es:

$$\bar{x} = \frac{\sum_{i=1}^n P_{mi} \cdot f_i}{\sum_{i=1}^n f_i} = \frac{166}{32} \approx 5,1875$$

Considerando que la media aritmética de esta distribución es de 5,1875; el cálculo de la variancia y desviación típica se efectúa de la siguiente manera:

x_i	f_i	P_{mi}	$P_{mi} - \bar{x}$	$(P_{mi} - \bar{x})^2$	$(P_{mi} - \bar{x})^2 \cdot f_i$
0,5 – 2,5	3	1,5	-3,6875	13,5976	40,7928
2,5 – 4,5	9	3,5	-1,6875	2,8477	25,6293
4,5 – 6,5	13	5,5	0,3125	0,0977	1,2701
6,5 – 8,5	4	7,5	2,3125	5,3477	21,3908
8,5 – 10,5	3	9,5	4,3125	18,5977	55,7931
Total	32	–	–	–	144,8761

La variancia es:

$$\sigma^2 = \frac{\sum_{i=1}^n f_i \cdot (P_{mi} - \bar{x})^2}{\sum_{i=1}^n f_i} = \frac{144,8761}{32} \approx 4,5274$$

La desviación típica es:

$$\sigma = \sqrt{4,5274} \approx 2,1278$$

Por lo tanto, la desviación típica de la clase A cuando los datos están agrupados en intervalos es 2,1278.



Coeficiente de variación

Frecuentemente se presenta el problema de comparar la dispersión de dos o más distribuciones. Por ejemplo, se quiere saber si la variabilidad de la temperatura de un lugar es mayor o menor que la del otro; si los sueldos varían más en un grupo de empleados que en otro, etc. En todos estos casos de comparación de dispersiones, ésta puede hacerse utilizando la desviación standar, si las variables comparadas tienen sus medias aritméticas iguales o aproximadamente iguales y si dichas variables vienen expresadas en la misma unidad de medida. Así, una desviación standar de 2,8 m no tiene el mismo significado cuando se trata de un conjunto de datos medidos en km, que cuando dichos datos se han medido en metros. Para evitar este inconveniente, se introducen parámetros que miden la desviación relativa y que se expresan mediante números carentes de dimensión. Este parámetro recibe el nombre de **coeficiente de variación**.

Se define al **coeficiente de variación** como el cociente entre la desviación típica y la media aritmética, que suele expresarse en tanto por ciento; no es otra cosa que la desviación estándar expresada como porcentaje de la media aritmética.

$$C_v = \frac{\sigma}{\bar{x}} \cdot 100$$

Este parámetro, al carecer de dimensiones, permite comparar el comportamiento de datos estadísticos medidos en unidades distintas dentro de una magnitud (metros o kilómetros) e, incluso, cuando se trata de magnitudes distintas (litros, kilogramos, años, etc).

La distribución que tiene menor dispersión es aquella que presenta un coeficiente de variación menor.

Distribución normal

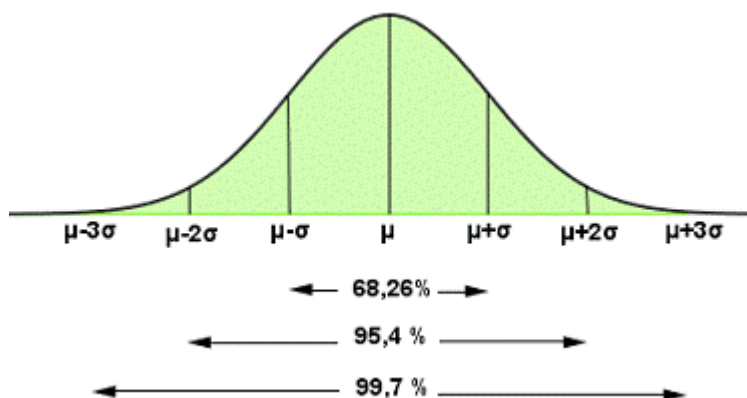
Una de las distribuciones más importantes en Estadística es la denominada **distribución normal**. Fue introducida por Carl Gauss a principios del siglo XIX en su estudio de los errores de medida. Desde entonces, se ha utilizado como modelo en multitud de variables (peso, altura, calificaciones...), en cuya distribución los valores más usuales se agrupan en torno a uno central y los valores extremos son escasos.

La distribución normal es campaniforme y simétrica. Por esta razón, es que en este tipo de distribuciones la media aritmética, la mediana y el modo coinciden. Además, se cumplen las siguientes relaciones:

- En el intervalo $]\bar{x} - \sigma ; \bar{x} + \sigma [$ se halla el 68,26% de los datos.
- En el intervalo $]\bar{x} - 2\sigma ; \bar{x} + 2\sigma [$ se halla el 95,4% de los datos.
- En el intervalo $]\bar{x} - 3\sigma ; \bar{x} + 3\sigma [$ se halla el 99,7% de los datos.

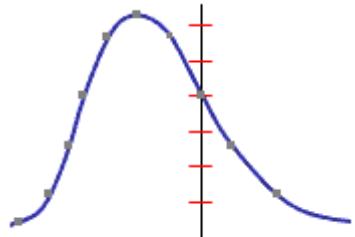
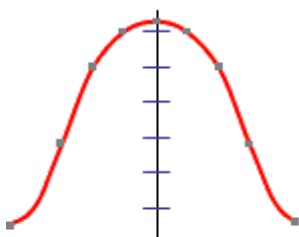
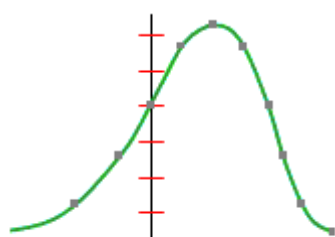
Estas relaciones sirven también para averiguar si los datos de un conjunto estadístico siguen una distribución normal. Para ello, se calcula su media y su desviación típica y se comprueba si en los correspondientes intervalos hay un porcentaje de datos semejantes al que afirma esta propiedad.

Esta afirmación también es cierta para las distribuciones simétricas o moderadamente asimétricas. Cualquiera de estos intervalos es a su vez una medida de dispersión, ya que si dicho intervalo es muy pequeño, implica que las observaciones están muy concentradas y si es muy grande es que están muy dispersos.



Asimetría

En las distribuciones campaniformes simétricas, la media aritmética y el modo coinciden. Pero si no es simétrica estos dos promedios no coincidirán porque uno corresponde al centro de gravedad de la figura y el otro a la máxima ordenada. Por lo tanto, la diferencia entre la media aritmética y el modo es una medida de la asimetría expresada en las mismas unidades que la variable.

Distribución asimétrica a la derecha	Distribución simétrica	Distribución asimétrica a la izquierda
$\bar{x} - M_0 > 0$  <p style="text-align: center;">Eje de simetría</p>	$\bar{x} - M_0 = 0$  <p style="text-align: center;">Eje de simetría</p>	$\bar{x} - M_0 < 0$  <p style="text-align: center;">Eje de simetría</p>



Actividades

1) Se miden las pulsaciones de 40 operarios para saber si el tipo de trabajo realizado por ellos afecta a la frecuencia cardíaca. Los resultados que se obtienen son los siguientes:

81	76	82	81	79	78	80	81	82	79
74	79	75	81	80	84	83	76	83	82
78	75	82	79	78	82	84	79	81	80
80	80	77	80	81	78	80	80	77	76

a) Construir la tabla de frecuencias y calcular la frecuencia relativa, porcentual y acumulada.

b) ¿Cuál es el mayor número de pulsaciones?

c) ¿Qué porcentaje de operarios alcanzó 80 o menos pulsaciones?

d) ¿Cuál fue el número de pulsaciones que alcanzó la mayor frecuencia?

e) Realizar un gráfico de los datos y extraer conclusiones.

2) En una fábrica se hizo una investigación que consistía en observar el número de tornillos que había en cada uno de 65 lotes que se habían comprado a un precio menor que el del mercado. Se obtuvo la siguiente información:

0	3	2	4	5	4	3	4	7	8	5	3	4
1	5	3	4	2	4	2	9	4	6	4	3	4
5	1	2	4	3	2	0	7	6	2	3	5	4
0	3	5	2	2	4	8	6	1	4	3	5	3
4	1	6	3	4	3	6	2	3	4	1	3	2

a) Construir la tabla de frecuencias y calcular la frecuencia relativa, porcentual y acumulada.

b) ¿Qué porcentaje de lotes tuvo menos de 4 tornillos defectuosos?

c) ¿Cuál fue el número de tornillos defectuosos que alcanzó la menor frecuencia?

d) Realizar un gráfico de los datos y extraer conclusiones.

3) Se encuestaron a 30 empleados para que manifestaran el grado de acuerdo o desacuerdo con respecto a la política empleada por la empresa en la que trabajan. A continuación, se muestran los resultados obtenidos:

Indeciso	Muy en desacuerdo	En desacuerdo	Indeciso	Muy de acuerdo
De acuerdo	Muy de acuerdo	Indeciso	En desacuerdo	Indeciso
En desacuerdo	Indeciso	De acuerdo	Muy en desacuerdo	De acuerdo
Indeciso	De acuerdo	Indeciso	Indeciso	Indeciso
Muy de acuerdo	En desacuerdo	En desacuerdo	De acuerdo	En desacuerdo
Muy en desacuerdo	De acuerdo	Indeciso	De acuerdo	Indeciso



- a) Construir la tabla de frecuencias y calcular la frecuencia relativa y porcentual.
- b) ¿Qué porcentaje de personas estuvo de acuerdo con la política implementada por la empresa?
- c) ¿Cuál fue el grado de acuerdo o desacuerdo que alcanzó la mayor frecuencia?
- d) Realizar un gráfico de los datos y extraer conclusiones.

4) Los siguientes datos corresponden a las medidas de los diámetros en centímetros de una muestra de 60 piezas manufacturadas por una fábrica:

1,738	1,738	1,735	1,744	1,725	1,740	1,735	1,742	1,731	1,732
1,735	1,735	1,738	1,735	1,732	1,737	1,732	1,734	1,730	1,734
1,736	1,736	1,735	1,743	1,727	1,733	1,737	1,742	1,741	1,732
1,739	1,735	1,730	1,726	1,734	1,736	1,734	1,739	1,737	1,741
1,728	1,729	1,739	1,724	1,732	1,733	1,736	1,735	1,736	1,746
1,733	1,731	1,727	1,745	1,729	1,730	1,728	1,736	1,740	1,740

- a) Construir la tabla de frecuencias utilizando intervalos de clases.
- b) Calcular la frecuencia relativa, porcentual y acumulada.
- c) Confeccionar el histograma y el polígono de frecuencia.
- d) Formular tres preguntas en base a los datos y redactar la conclusión correspondiente.

5) Se han pesado 40 piezas metálicas. Los resultados de las pesadas, expresados en gramos son:

64,1	66,6	67,3	66,1	64,4	65,8	65,0	63,1
68,8	66,4	67,0	65,3	65,7	64,1	64,5	64,3
65,0	66,3	66,7	68,5	61,5	64,4	64,6	63,5
66,9	64,0	64,2	64,0	65,3	63,0	65,4	63,2
66,4	65,1	65,7	66,8	63,9	63,1	63,0	65,5

- a) Construir la tabla de frecuencias utilizando intervalos de clases.
 - b) Calcular la frecuencia relativa, porcentual y acumulada.
 - c) Confeccionar el histograma y el polígono de frecuencia.
 - d) Formular tres preguntas en base a los datos y redactar la conclusión correspondiente.
- 6) Teniendo en cuenta la distribución de frecuencias obtenida en el **Ejercicio 1**, calcular:
- a) Media aritmética, mediana y modo.
 - b) Primer y tercer cuartil.
 - c) Segundo y noveno decil.
 - d) Analizar los resultados y extraer conclusiones.



7) Calcular el parámetro de tendencia central que considere adecuado en la distribución de frecuencias obtenida en el **Ejercicio 3**.

8) Teniendo en cuenta la distribución de frecuencias obtenida en el **Ejercicio 4**, calcular:

- a) Media aritmética, mediana y modo.
- b) Primer y tercer cuartil.
- c) Segundo y octavo decil.
- d) Analizar los resultados y extraer conclusiones.

9) Una fábrica compuesta por 1800 obreros, se ha dividido en dos unidades estratégicas. La siguiente tabla muestra la distribución de sueldos en las dos unidades estratégicas.

	<i>Planta Norte</i>	<i>Planta Sur</i>
<i>Sueldo (medido en \$)</i>	<i>Número de obreros</i>	<i>Número de obreros</i>
3500	120	100
4000	400	200
4500	160	480
5000	80	100
5500	32	80
6000	8	40

a) Calcular la media aritmética, mediana, modo y desviación estándar de los sueldos en cada planta.

b) ¿Qué porcentaje de empleados, en cada planta, cobra más de \$5000?

c) ¿Cuál de las dos empresas tiene mayor homogeneidad salarial? ¿Por qué?

d) Calcular los intervalos $]\bar{x} - \sigma; \bar{x} + \sigma[$, $]\bar{x} - 2\sigma; \bar{x} + 2\sigma[$ y $]\bar{x} - 3\sigma; \bar{x} + 3\sigma[$ para ver si la distribución se aproxima a la normal.

e) Representar gráficamente los datos y extraer conclusiones.

10) La siguiente tabla muestra la distribución de frecuencias de puntuaciones de un test realizado a 120 empleados.

<i>Puntuaciones</i>	<i>Número de empleados</i>
30,5 – 40,5	9
40,5 – 50,5	32
50,5 – 60,5	43
60,5 – 70,5	21
70,5 – 80,5	11
80,5 – 90,5	3
90,5 – 100,5	1



- a)** Hallar las medidas de tendencia central y dispersión.
- b)** Calcular los cuartiles de la distribución e interpretar resultados.
- c)** Determinar la puntuación más alta alcanzada por el 20% más bajo del grupo.
- d)** Calcular los intervalos $]\bar{x} - \sigma; \bar{x} + \sigma [$, $]\bar{x} - 2\sigma; \bar{x} + 2\sigma [$ y $]\bar{x} - 3\sigma; \bar{x} + 3\sigma [$ para ver si la distribución se aproxima a la normal.
- e)** Confeccionar el histograma y el polígono de frecuencias.
- f)** Extraer conclusiones.